USGS
science for a changing world

Data Release Use Case Team: Status of effort and issues identified

September 2013

Team Leads: Fran Lightsom and Viv Hutchison

Presenter: Keith Kirk

U.S. Department of the Interior
U.S. Geological Survey

Members of the CDI Use Case Team

- Fran Lightsom (co-lead), Natural Hazards (Woods Hole, MA)
- Viv Hutchison (co-lead), Core Science Systems (Denver, CO)
- John Faundeen, Climate and Land Use Change (Sioux Falls, SD)
- Greg Gunther, Energy and Minerals (Denver, CO)
- Keith Kirk, Office of Science Quality and Integrity (Santa Cruz, CA)
- Greg Miller, Natural Hazards (St. Petersburg, FL)
- Andrea Ostroff, Core Science Systems (Reston, VA)
- Carolyn Reid, Office of Science Quality and Integrity (Reston, VA)

- Facilitator: Peter Fox, Rensselaer Polytechnic Institute (Troy, NY)

Cross-Mission Area Representation

USGS

Background: Use Case Team

- **Original purpose:** to develop a process, based on current policies and workflows, that enables USGS employees to determine if a particular set of data is approved for release

- Convened early 2012 through the USGS Community for Data Integration (CDI) Data Management Working Group
  - Met face-to-face in Reston VA in April 2012
  - Weekly phone meetings ever since *(whew!)*

USGS

Major Challenges Initially Identified:

- **Lack of:**
  - bureau-wide understanding of policies and procedures for releasing data
  - bureau-wide understanding of distinctions between publishing data in a USGS series report versus other means of data release
  - attention to data preservation within Fundamental Science Practices
  - Explanation of differences between peer review and data review is not reflected in current policy
- **Resistance to:**
  - metadata creation along with inconsistent or absent treatment of metadata in the release process

USGS

Connections to External Drivers

- **Open Government Initiatives supporting broader public access to Federal and Federally-supported data and information**
  - *the Use Case Team thinking was ahead of recent directives, now we have the opportunity to leverage these directives to facilitate positive change in the USGS*

Policy	Type	Date Issued
Transparency and Open Government	Presedential Memorandum	1/21/2009
Open Government Directive	OMB Memorandum M–10–06	12/8/2009
Digital Government: Building a 21st Century Platform to Better Serve the American People	Federal CIO Strategy Document	5/23/2012
Managing Government Records Directive	OMB-NARA Memorandum M–12–18	8/24/2012
Increasing Access to the Results of Federally Funded Scientific Research	OSTP Memorandum	2/22/2013
Making Open and Machine Readable the New Default for Government Information	Executive Order 13642	5/9/2013
Open Data Policy-Managing Information as an Asset	OMB Memorandum M–13-13	5/9/2013

Web Release Use Case: Diagrams

USGS

Data Release via Web: Assumptions

To build the Use Case, some assumptions were made to start:

- **Data is not interpretive**
- **USGS data product is assumed to be non-proprietary, and non-sensitive**
- **Science Center web sites can host data**

Data Release via Web: Pre-Conditions

To build the Use Case, we noted existing "pre-conditions" before someone would use this workflow:

- **USGS data exists and is available to the Author**
- **Data has not been previously released nor is part of a national collection or other 'approved' dissemination**
- **Data are not pre-decisional (SM 502.5)**
- **A web site that is appropriate for release of the data exists or can be created and is available to the Author.**
- **Data and datasets are non-interpretive and therefore require a *data* quality review not a traditional peer review.**
  - **Peer review is appropriate for interpretive information products (scholarly publications).**

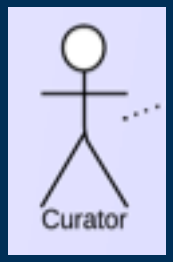
USGS

Actors

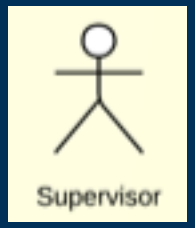
- **A use case establishes the relevant "actors"…what roles are involved?**

creates and revises product, initiates product approval request, prepares metadata for product

reviews product for scientific quality



manages product preservation, generally this is the originating Science Center or NatWeb



reviews for conformity with FSP policies and processes

approves product for release

oversees application of digital object identifier for data and metadata

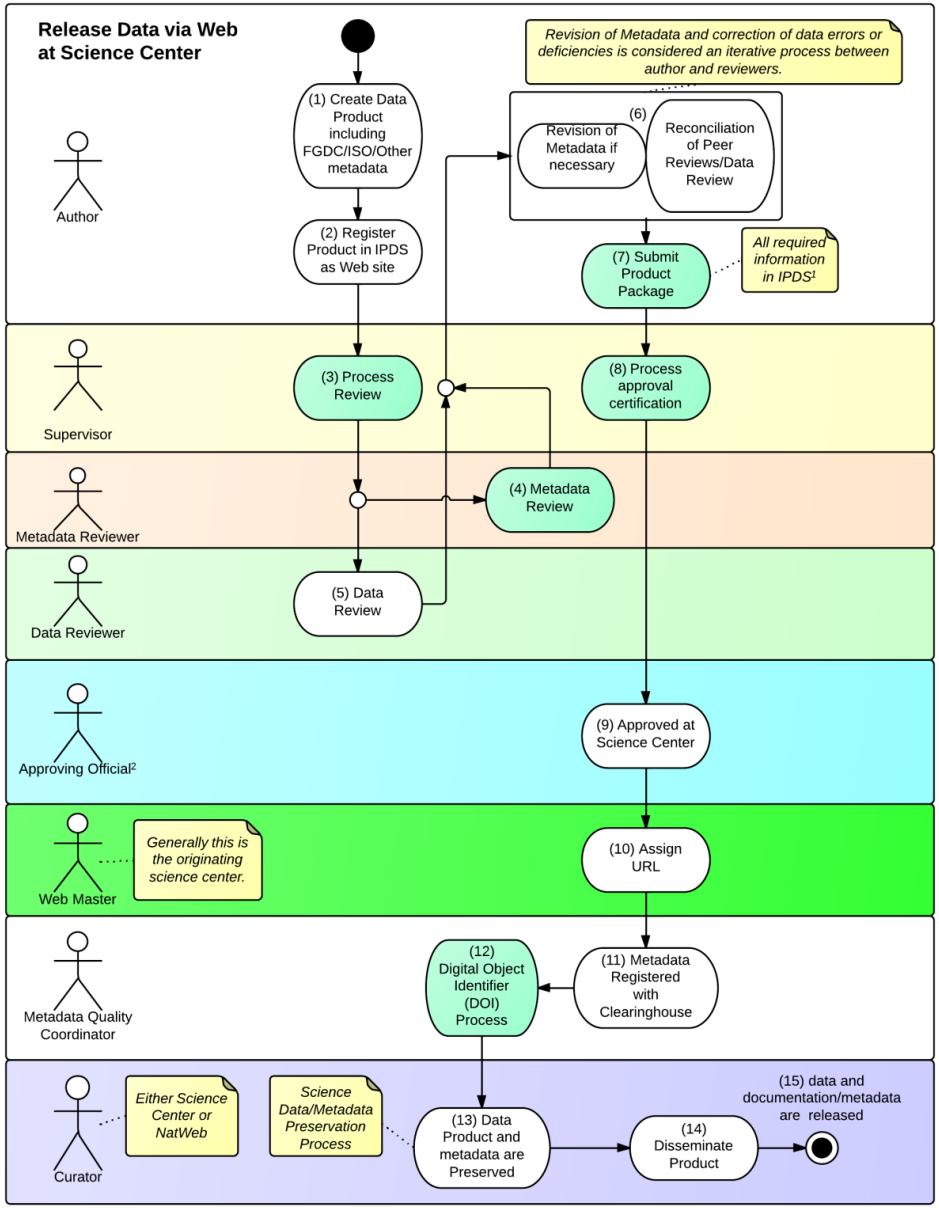
reviews metadata for accuracy and conformance with standards

generally this is the originating Science Center

Note: *One person may act in multiple roles*

Data Release via Web:

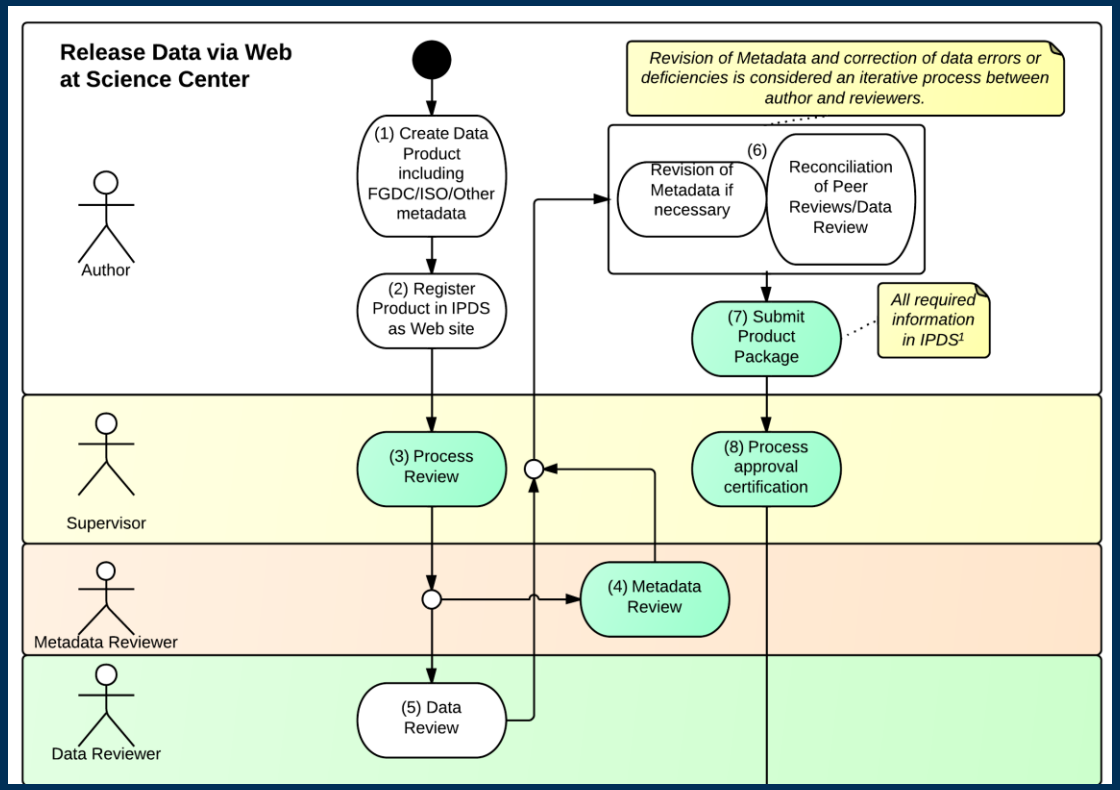
Workflow Diagram Overview



1. IPDS is the Bureau **I**nformation **P**roduct **D**ata **S**ystem. Refer to URL: http://internal.usgs.gov/publishing/ipds.html
2. Under FSP approval of non-interpretive information products is delegated to the Science Center Director

USGS

Use Case: The details



Link to PDF:

What did we learn?
Use Case Recommendations

1. **Implement Digital Object Identifiers** or DOIs, to enable identification of both data that are approved for release and the associated publications.
2. Search capabilities for **publication catalogs that link to data**. (Possible with new Pubs Warehouse, ScienceBase2)
3. Implement USGS online data services that **provide data in a useful format**.
4. **Training** for providing data in useful formats on web.
5. **Guidance** for choice of data reviewers.
6. **Guidance** on responsibilities related to data review.
7. Hold data until data is reviewed and metadata is appropriate.
8. Require a **DOI in metadata record** to identify data that is approved for release.
9. **Enforce metadata review** for data that is released.
10. Education and resources for metadata reviewers.
11. Guidance for determining interpretive content.
12. **Policies and guidance for data preservation** (FSPAC working on this).
13. Develop approved **online repositories** that will preserve USGS data and information and ensure that they can be found and used in the future.
14. **Mandatory** training on data release (i.e., DOI Learn).
15. Provide information about new policies and processes for web release of data.
16. Establish **RGE credit** for high-quality data release.
17. Establish **data citation standards** so scientists get credit.

Data Release Use Case Team: Recommendations

April 2, 2013

Introduction

In March 2012, the Data Release Use Case Team convened through the USGS Community for Data Integration (CDI) Data Management Working Group to develop a procedure that enables USGS employees to determine if a particular set of data is approved for release.

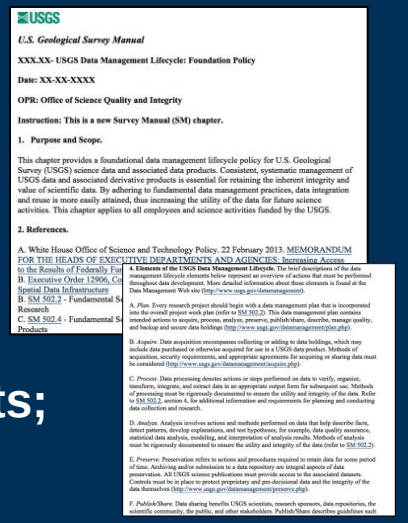
The team is comprised of representatives from various mission areas, including Climate and Land Use Change, Core Science Systems, Energy and Minerals and Environmental Health, and Natural Hazards, and the Office of Science Quality and Integrity. The skill set of the nine-member team includes data experts, approving officials, software engineers, and experts in information modeling and USGS policy (refer to Appendix A for the members list). The team used a semantic development (Use Case) approach, facilitated by Dr. Peter Fox of Rensselaer Polytechnic Institute (RPI). More information on the team's activities is available on the CDI Confluence site at https://my.usgs.gov/confluence/display/cdi/Data+Release+Use+Case+Team.

The team's work involved developing use case diagrams that depict the current and or recommended flow of the review and approval process for various data release scenarios. The first (referred as

USGS

Connecting the Dots in USGS policy

- **In Progress**
  (via the "Data Mgmt Policy Group" & FSPAC)
  - Data Management Overarching Policy
    - Based on USGS Data Lifecycle
  - Metadata for Datasets and Information Products
    - Requires FGDC endorsed standards for data products; use of IPDS metadata for publications
  - Release of Software
  - Release of Data
    - Requires application of Digital Object Identifiers (DOIs) for all data (to connect to publications)
- **Other policies needed:**
  - Data citation standards
  - RGE recognition so that scientists get credit for releasing data
  - Data preservation standards


USGS

U.S. Geological Survey Manual

XXX.XX- USGS Data Management Lifecycle: Foundation Policy

Date: XX-XX-XXXX

OPR: Office of Science Quality and Integrity

Instruction: This is a new Survey Manual (SM) chapter.

1. Purpose and Scope.

This chapter provides a foundational data management lifecycle policy for U.S. Geological Survey (USGS) science data and associated data products. Consistent, systematic management of USGS data and associated derivative products is essential for retaining the inherent integrity and value of scientific data. By adhering to fundamental data management practices, data integration and reuse is more easily attained, thus increasing the utility of the data for future science activities. This chapter applies to all employees and science activities funded by the USGS.

2. References.

A. White House Office of Science and Technology Policy. 22 February 2013. MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES: Increasing Access to the Results of Federally Fu

B. Executive Order 12906, C Spatial Data Infrastructure

B. SM 502.2 - Fundamental S Research

C. SM 502.4 - Fundamental S Products

4. Elements of the USGS Data Management Lifecycle. The brief descriptions of the data management lifecycle elements below represent an overview of actions that must be performed throughout data development. More detailed information about these elements is found at the Data Management Web site (http://www.usgs.gov/datamanagement).

A. Plan. Every research project should begin with a data management plan that is incorporated into the overall project work plan (refer to SM 502.2). This data management plan contains intended actions to acquire, process, analyze, preserve, publish/share, describe, manage quality, and backup and secure data holdings (http://www.usgs.gov/datamanagement/plan.php).

B. Acquire. Data acquisition encompasses collecting or adding to data holdings, which may include data purchased or otherwise acquired for use in a USGS data product. Methods of acquisition, security requirements, and appropriate agreements for acquiring or sharing data must be considered (http://www.usgs.gov/datamanagement/acquire.php).

C. Process. Data processing denotes actions or steps performed on data to verify, organize, transform, integrate, and extract data in an appropriate output form for subsequent use. Methods of processing must be rigorously documented to ensure the utility and integrity of the data. Refer to SM 502.2, section 4, for additional information and requirements for planning and conducting data collection and research.

D. Analyze. Analysis involves actions and methods performed on data that help describe facts, detect patterns, develop explanations, and test hypotheses; for example, data quality assurance, statistical data analysis, modeling, and interpretation of analysis results. Methods of analysis must be rigorously documented to ensure the utility and integrity of the data (refer to SM 502.2).

E. Preserve. Preservation refers to actions and procedures required to retain data for some period of time. Archiving and/or submission to a data repository are integral aspects of data preservation. All USGS science publications must provide access to the associated datasets. Controls must be in place to protect proprietary and pre-decisional data and the integrity of the data themselves (http://www.usgs.gov/datamanagement/preserve.php).

F. Publish/Share. Data sharing benefits USGS scientists, research sponsors, data repositories, the scientific community, the public, and other stakeholders. Publish/Share describes guidelines such

Build It, Test It

- **Washington Water Science Center**

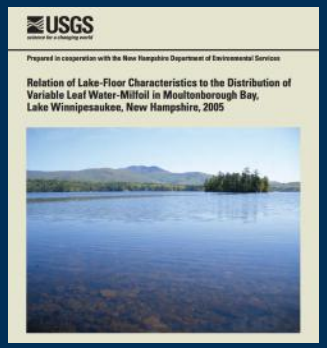
  (First data set with DOIs released 6/10/2013 in response to litigation associated with Elwha Dam Removal)**.** Data released initially to meet immediate need to provide data in support of litigation against Interior related to consequences of Elwha River dam removal effort. Ongoing data collected daily is added to site after QA/QC

- **Pacific Science Center: Santa Cruz**

  Rich and VERY large data set. Not appropriate for Data Series owing to data type. Large data stacks using Open Access. YouTube and Picasso leveraged as data repositories. Separate YouTube channel approved by OCAP for this data. Approved using new IPDS. Currently served on non-USGS site will be moved to USGS site as resources allow. This will be easily facilitated by DOI metadata. Metadata for all video and images per OCAP requirements.

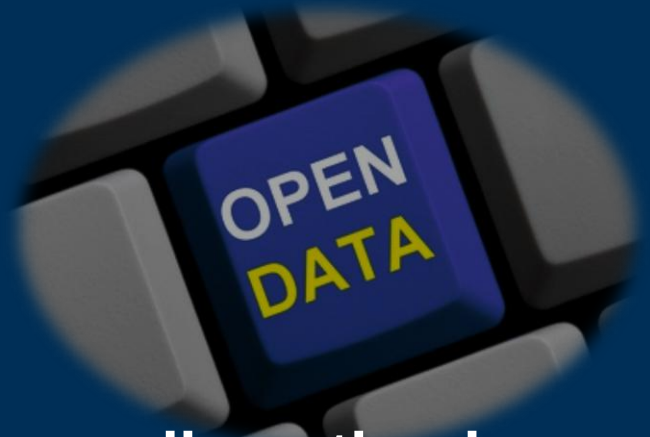
- **Woods Hole**

  Data were used in figures in a 2007 Scientific Investigations Report (SIR). The data were not published at the time but now requests for data are numerous. Data will be released outside of USGS series to facilitate need for machine readable data. Data will go online with a metadata record after an FSP-compliant review and approval process. Data will probably be part of the existing Woods Hole Science Center data library Website.



USGS

Benefits of Web Release for USGS Data



- **Eliminates costs associated with producing and distributing media**
  - **uses existing IT resources**

- **Allows science center directors to <u>use discretion in designating materials for editorial review</u>**
  - more efficient use of the editors' time and a higher rate of information product publication.
- **Open Data Compliance: Data online can be hosted in <u>formats that allow re-use</u>**
  - Moving away from a pdf format to open data formats will greatly enhance our ability to share data and make our data more interoperable
  - Compliance with OSTP memo and other directives.
- **Helps meet OSTP & OMB Directives related to open access, data release & Machine Readability**

USGS

Next Steps for Data Release Team

- Seek review and approval of Use Cases:
  - by broader groups (CDI DMWG, SDCN, etc)
  - by Alan and Kevin and USGS ELT
- Post Use Cases on USGS Data Management website
- Develop:
  - more use cases
  - more recommendations
  - education components
  - Ensure relevant policies and processes are developed (i.e., data release policy, digital object identifier guidance, etc)
- Get the word out – make processes clearer and easier

USGS

USGS

Thank you!

Questions and Comments

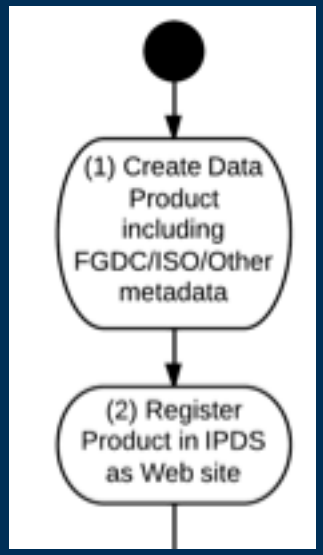
Oh! Did you say you want to review use case materials? ☺

- https://my.usgs.gov/confluence/display/cdi/Web+Release+Use+Case
- send comments to Fran (flightsom@usgs.gov) and Viv (vhutchison@usgs.gov)

USGS

Process Details

creates and revises product, initiates product approval request, prepares metadata for product

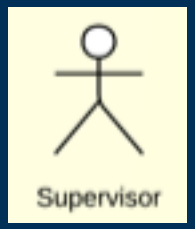


Author develops Product:

- Identifies the appropriate Web site (or Web data service)
- Makes available for review:
  - one or more clean data sets in appropriate format;
  - one or more metadata records;
  - any additional descriptive materials needed to ensure the data are discoverable and useful;
  - draft Web site

USGS

Process Details



reviews for conformity with FSP policies and processes

Process Review:

- Supervisor verifies appropriateness of choice of Web release;
- ensures the draft Web site meets standards for completeness,
- verifies choice of data reviewers, metadata reviewers;
- ensures compliance with relevant policy requirements
- gives author permission to provide draft web site to reviewers.

USGS

Process Details

reviews metadata for accuracy and conformance with standards

Metadata Review:

•Produces written report and returns this report to the author.

•Metadata Review Process:

- Check compliance using a recommended metadata validation tool.
- Perform metadata quality checks:
  - Check that the metadata matches the data
  - Check that data field names and values are defined and consistent with information in entity/attribute section of metadata record
  - Check that bounding coordinates match location keywords
  - Check temporary on-line linkage to data exists (this link(s) will change when final DOI is assigned)
  - Check that information about processing steps, methodology, lineage are included in the record.
  - Does the metadata provide robust information about how to use the data files – access instructions, software requirements, data models, definitions of terms, and so on?

USGS

Process Details

reviews product for scientific quality

Data Review:

- Produces written report to the author with a recommendation that Product be released and a list of any recommended changes.

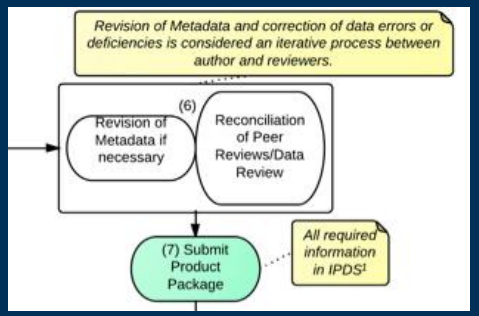
Data review may include the following:

- Is data format reasonable for public distribution (released using common standards)?
- Are data values reasonable? Are they in a valid range for that measurement, do they display any expected seasonal or daily trends, is there consistency between adjacent or otherwise related datasets?
- Can data be used by appropriate analysis and visualization tools?
- Does the metadata match the data?
- Are the data attributes listed in the metadata in agreement with the data?
- Are the techniques and methods scientifically sound and well described? Could a knowledgeable scientist or technician recreate the final data set from the descriptions? Can this information be easily found and used?
- Does the Product as a whole, through its design or documentation, provide enough information that the data and metadata can be easily found and used

USGS

Process Details

creates and revises product, initiates product approval request, prepares metadata for product



- Revision Process:
  - Author iteratively revises Product in response to and in cooperation with metadata reviewer and data reviewer comments and documents reconciliation and responses.

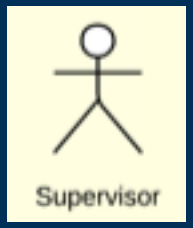
- Submit Package:
  - Author places all relevant materials in IPDS document vault as verification the review/reconciliation of the product took place and notifies Supervisor.

These materials include:

- A link to the draft Web site which has been reviewed and revised.
- Names of data reviewers and metadata reviewer, and evidence that they agree with revised form of the data and metadata.
- Reports from data and metadata reviewers, annotated by author to indicate changes made to data and metadata in response to reviews.
- Information about versioning, if the data is a first version that must be clearly indicated.
- URLs, DOIs, or bibliographic citations of publications that are related to the data.

USGS

Process Details



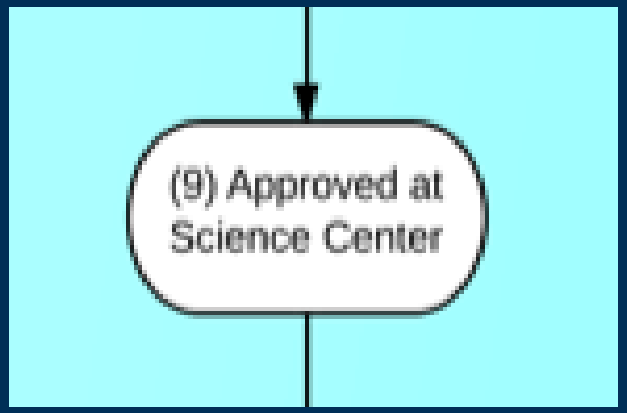
reviews for conformity with FSP policies and processes

Process Approval Certification:

- Supervisor verifies that the draft Web site is complete and responses/reconciliations are appropriate. Supervisor forwards request to local Approving Official (Science Center Director).

Process Details

approves product for release



Approval:

Non-interpretive data product is approved at science center or rejected.

(Science Center Director = Approving Official in case of non-interpretive data)

If rejected author is informed of need for additional changes.

Process Details

generally this is the originating Science Center

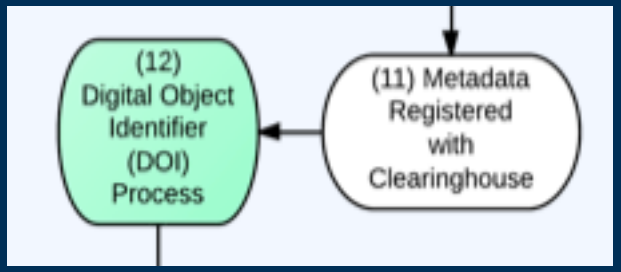
URL is assigned:

- **The URL assigned at this step represents the physical/actual URL**
  - **(EXAMPLE: http://energy.usgs.gov/data/dataset.zip)**
  - **URL is used as the de-referencable URL when obtaining the DOI**

USGS

Process Details

oversees application of digital object identifier for data and metadata

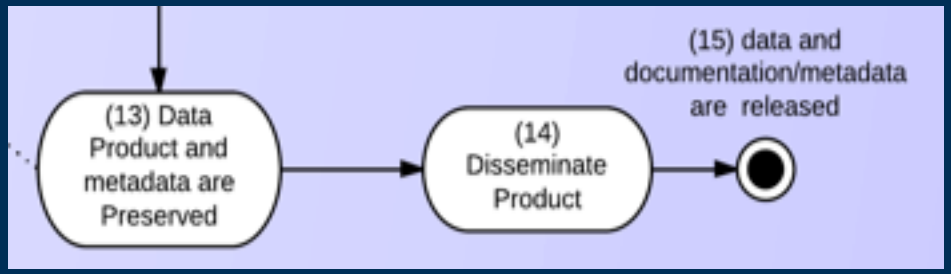


- **Digital Object Identifier obtained for data and populated into all relevant metadata, FGDC/ISO**
  - **Note: Updated metadata with new DOI (at least the DOI URL) must be passed back to author or responsible contact at Science Center**
- **Metadata Registered with Core Science Metadata Clearinghouse to make it available, according to Executive Order 12906 (1994)**

USGS

Process Details….final stretch

manages product preservation, generally this is the originating Science Center or NatWeb



Preservation:

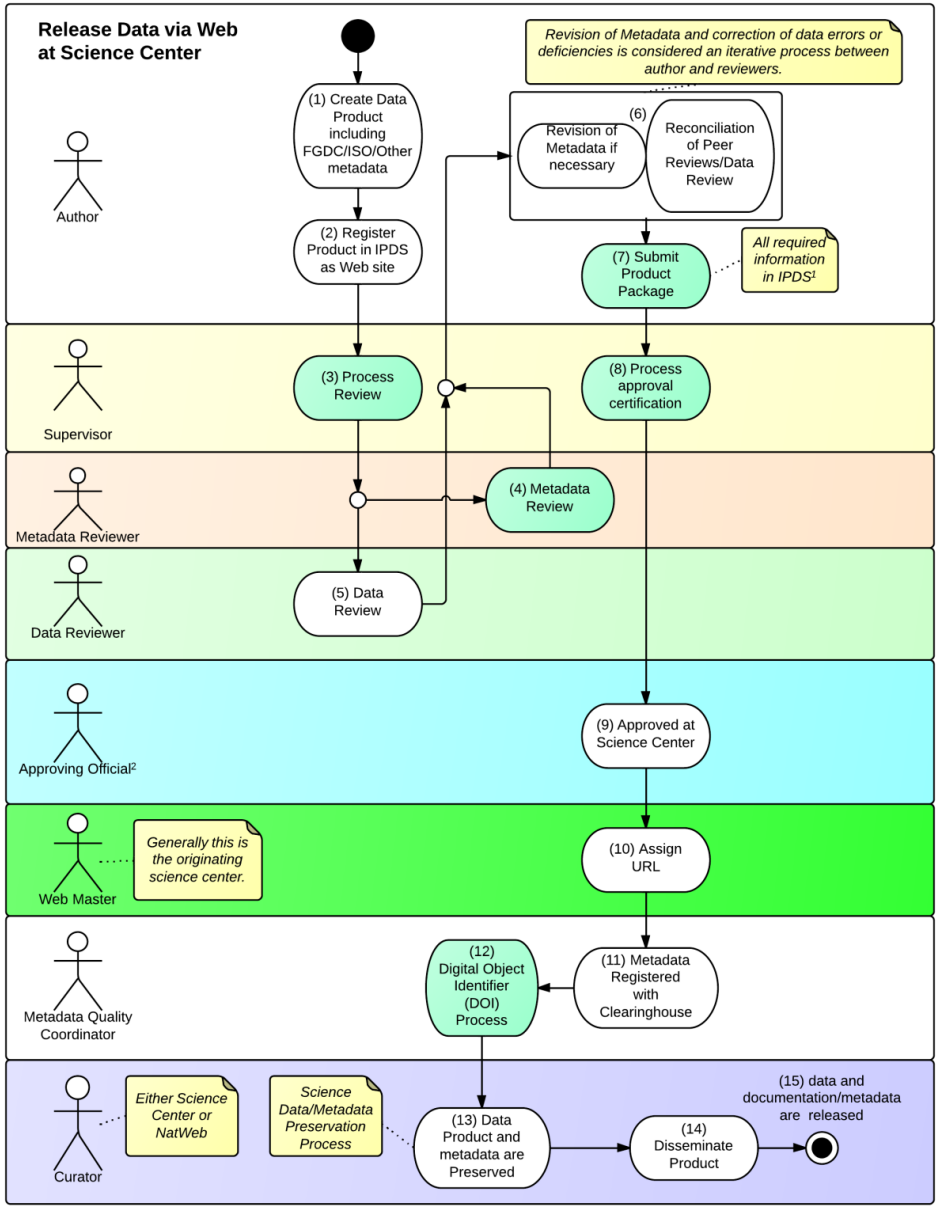
- **Data Product and metadata are preserved**
  - **If product is on NatWeb, then archive requirements are met**
  - **If product is being made available elsewhere, ensure both the metadata and the data are preserved according to standards.**

Data Release!

- **Approved Product is disseminated by Web release at Science Center.**

USGS

Data Release via Web



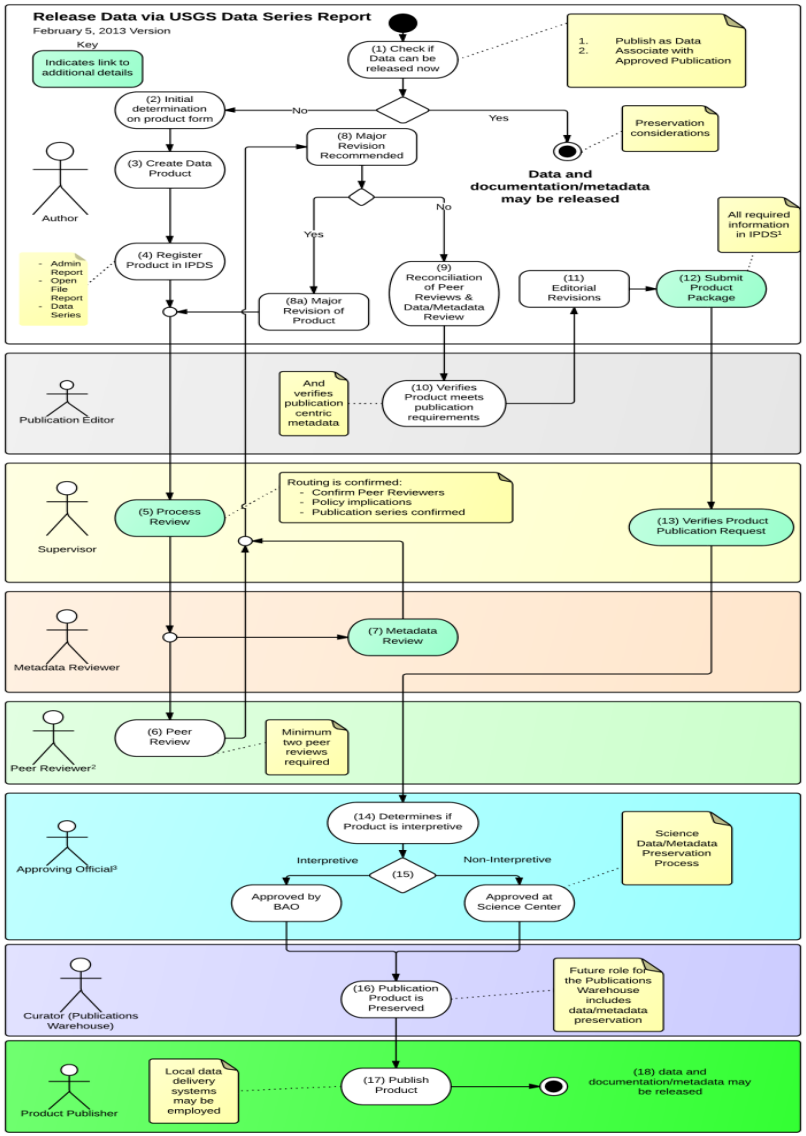
1. IPDS is the Bureau **I**nformation **P**roduct **D**ata **S**ystem. Refer to URL: http://internal.usgs.gov/publishing/ipds.html
2. Under FSP approval of non-interpretive information products is delegated to the Science Center Director

USGS

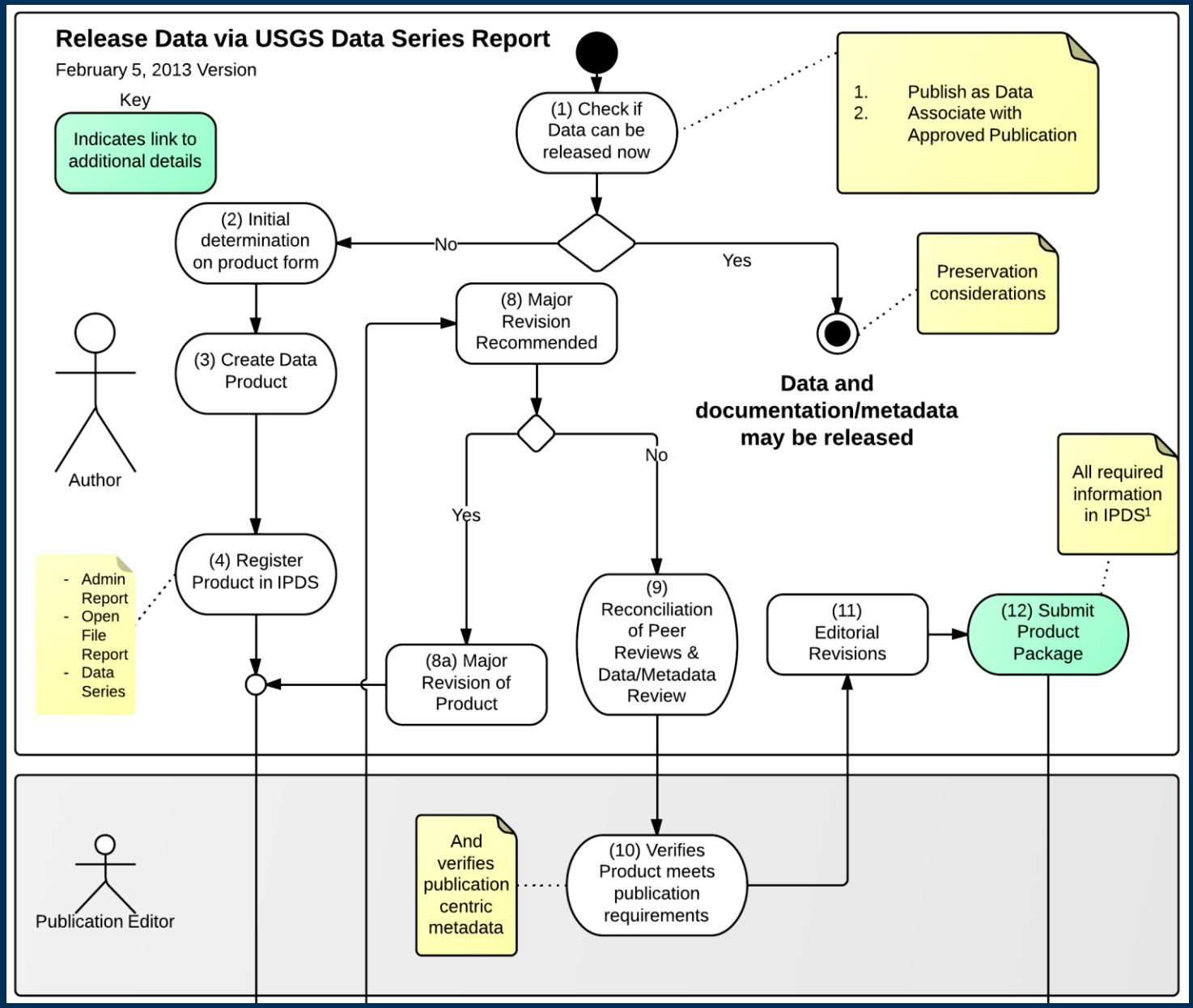
Additional Slides…

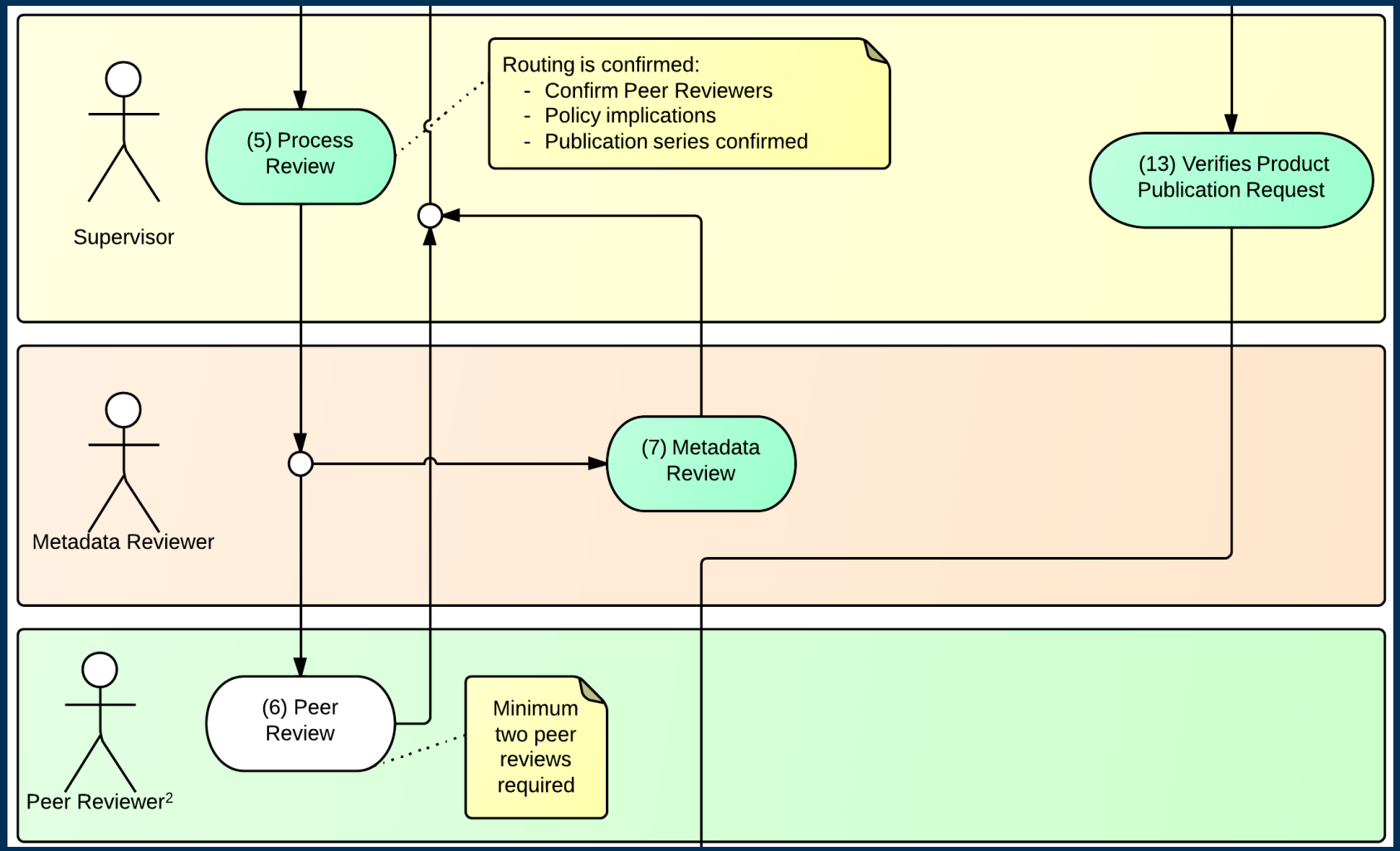
Data Release through USGS Series Publication

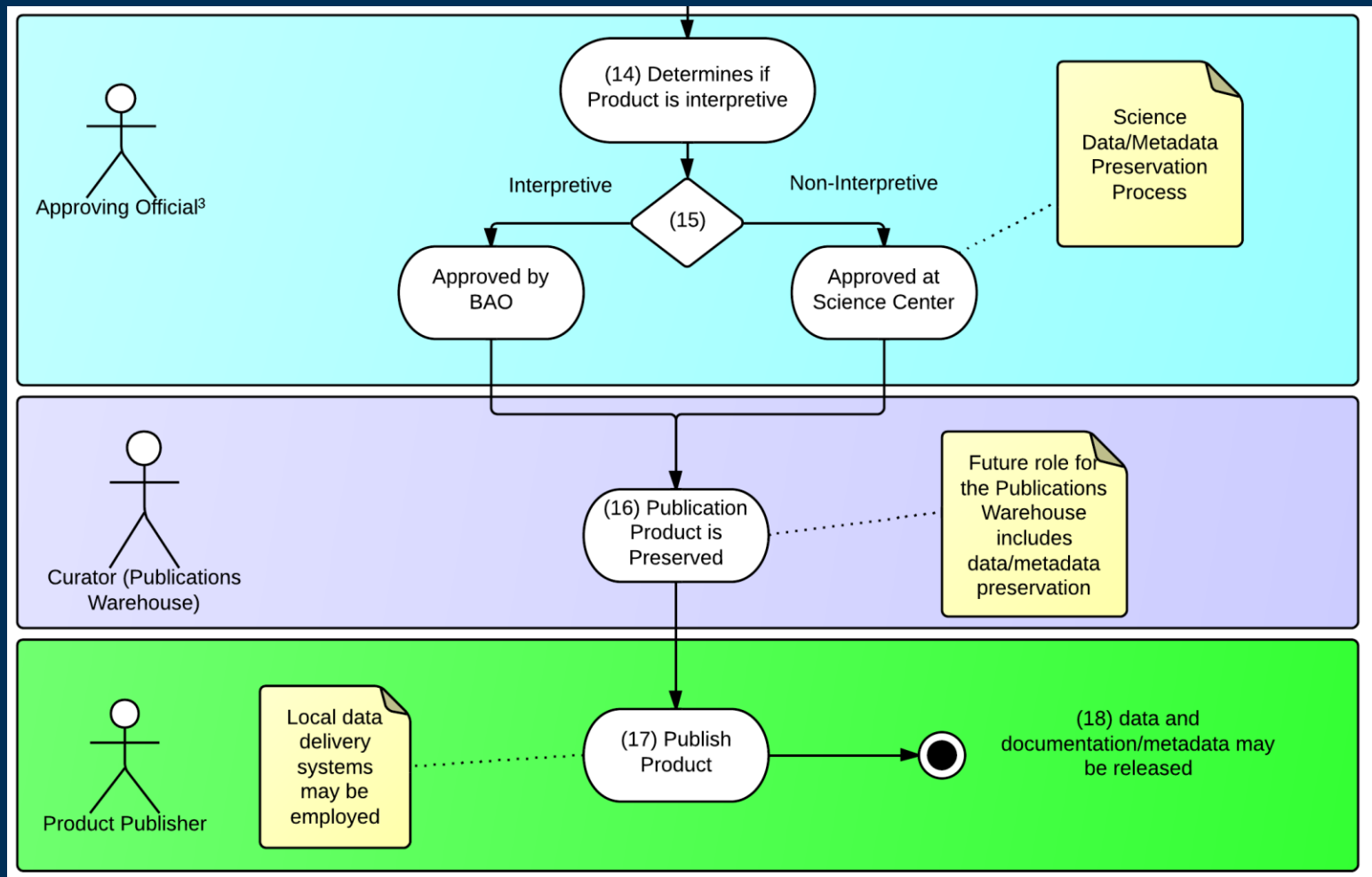
DATA SERIES USE CASE DIAGRAMS



1. IPDS is the Bureau **I**nformation **P**roduct **D**ata **S**ystem. Refer to URL: http://internal.usgs.gov/publishing/ipds.html
2. If reviewer(s) recommend that Product not be published, Author performs major revision to Product, possibly including extensive clean-up of data and/or metadata, and starts again at flow step 5
3. The Science Center Director or designee determines if product is interpretive or non-interpretive. If non-interpretive may approve. In addition, either the Science Center Chief or the BAO and "Kill" the process requiring additional information or a restart. Author performs major revision to Product, possibly including extensive clean-up of data and/or metadata, and starts again at basic flow step 2


Release Data via USGS Data Series Report
February 5, 2013 Version
Key
Indicates link to additional details
(1) Check if Data can be released now
1. Publish as Data
2. Associate with Approved Publication
(2) Initial determination on product form
No
Yes
Preservation considerations
(8) Major Revision Recommended
(3) Create Data Product
Author
Data and documentation/metadata may be released
No
Yes
All required information in IPDS[1]
(4) Register Product in IPDS
- Admin Report
- Open File Report
- Data Series
(9) Reconciliation of Peer Reviews & Data/Metadata Review
(11) Editorial Revisions
(12) Submit Product Package
(8a) Major Revision of Product
Publication Editor
And verifies publication centric metadata
(10) Verifies Product meets publication requirements


Supervisor
(5) Process Review
Routing is confirmed:
- Confirm Peer Reviewers
- Policy implications
- Publication series confirmed
(13) Verifies Product Publication Request
Metadata Reviewer
(7) Metadata Review
Peer Reviewer2
(6) Peer Review
Minimum two peer reviews required



1. IPDS is the Bureau **I**nformation **P**roduct **D**ata **S**ystem. Refer to URL: http://internal.usgs.gov/publishing/ipds.html
2. If reviewer(s) recommend that Product not be published, Author performs major revision to Product, possibly including extensive clean-up of data and/or metadata, and starts again at flow step 5
3. The Science Center Director or designee determines if product is interpretive or non-interpretive. If non-interpretive may approve. In addition, either the Science Center Chief or the BAO and "Kill" the process requiring additional information or a restart. Author performs major revision to Product, possibly including extensive clean-up of data and/or metadata, and starts again at basic flow step 2